

The authors of ‘Examining anthropogenic CO2 emission inventories over the Greater Tokyo Area using in situ observations and high-resolution atmospheric simulation’ compare CO2 mole fractions from different emission inventories at 4 observation sites in an urban area of Japan. The research is timely and could add to scientific advances made in understanding urban greenhouse gas emissions. However, in its current state, the research is not convincing enough to be published, and I encourage the authors to revise the manuscript. In general, I find the premise of the paper interesting, although the results and conclusion lack focus and could be improved, for this paper to become more interesting for a wider audience.

The main improvement to the paper, in my opinion, lies in the missing interpretation of the results. In the abstract, the authors state ‘The four observation sites represent different GTA subregions, providing complementary constraints on inventory spatial patterns’. However, such constraints on spatial patterns are never used in the manuscript. For example, the authors show that WRF-ODIAC over-estimates CO2 at YYG. This is despite ODIAC having the lowest total emissions over GTA. This suggests that the spatial distribution matters (as the authors state). What remains unanswered are the follow-up questions: What causes ODIAC to be too high here? Is there any process that ODIAC over-estimates? Is there an under-estimation at the other sites? What region influences YYG, and could that region specifically be constrained by these observations? I am missing such deepening insights throughout the manuscript.

Another issue I have with the current paper is the reliance on atmospheric transport. Urban atmospheric transport is notoriously hard, and the authors show consistent biases in the wind direction at most sites. The authors select only observations from 12-15 LST, which is common practice. However, given that observations and simulations of wind speed are available, a more sophisticated selection would make more sense. For example, only select observations during which WRF performed well in wind speeds and direction. I believe this would reduce the high simulated FF contributions of up to 400ppm. I think the manuscript would gain a lot from such an analysis. Beside these two main points, I have some general and detailed comments which I outline below.

1 General comments

It would be good to analyse differences in temporal disaggregation. Using different diurnal cycles might influence the results significantly, adding an interesting layer of uncertainty to the simulations. The biosphere might be very important, and I’m not convinced by the comparison to one single EC tower (where the authors say that VPRM has a flatter diurnal cycle than observed). From Fig. 5, it seems that the biosphere does not contribute strongly. However, a bias of 1 ppm (well within the biosphere contribution) would influence the results in Fig. 9 strongly. I would expect a more thorough analysis of these contributions, and their potential errors, for example by analysing dCO2 from different biosphere products.

The uncertainty in the background should be taken into account for the calculation of the ‘updated’ inventories. From the slope shown in Figure 9, $1 \text{ ppm} \pm 5 \text{ TgC/y}$. The reported 0.2 ppm bias (which I think is a low-end estimate) would thus result in $\pm 1 \text{ TgC/y}$, which is 5% of spread between inventories. The authors shortly mention the background in the discussion, but not sufficiently.

Methods are sprinkled throughout the results (for example the 2std filter on the simulated CO2

in Fig. 10 and the 12-15 LST filter). Please separate them.

L633: ‘provides a practical benchmark for improving urban CO₂ inventories’. I am not sure what the benchmark the authors refer to is. If they mean that different station locations can be used for quantification of different sectors / regions, I would urge the authors to also add such an analysis.

2 Detailed comments

In Figure 4, the downscaling of 3/5 models results in strange artifacts. I think a more fair comparison would be to regrid MOSAIC to the coarser resolution. Also the unit is not logarithmic, the colorbar is.

It is hard to see the main messages from Fig.6. Consider simpler figures

Figure 7 seems to miss the split between summer and winter PBLH.

L.251: Integrated over the year, NEE is small. But, with a strong diurnal and seasonal cycle. So on any given hour, it is probably in the same range as FF.

L.289. To analyse the uncertainty in the background, it might be good to analyse the mole fractions from other models (or explain why only CAMS was used).

L.473: The split between summer and winter is not shown.

L.510: The mean bias seems to be affected strongly by the time selection of the data. Starting an hour earlier or later seems (from Fig. 7) to result in a different bias, which would in turn result in a different ‘best estimate’ for the total emissions. It would be good to explore this uncertainty, although it might change with my previous comments.

L543: No red line is shown. It is not clear what ‘coloured bins show 2d histograms’ means, as only a scatter is shown. Are the 2std outliers also removed for regression calculation, or only for plotting purposes?

L559: ‘Overall, the inventories capture much of the temporal variability, but notable scatter remains’. The FF enhancement only explains up to 32% of the variance, I agree that this can be classified as ‘much’, but some quantification and nuance is required here.

L561: It is unclear how the uncertainties are calculated and propagated. Is it only the uncertainty on the slope? What about uncertainty due to site selection, background, biosphere and transport?

L605: These ‘complementary constraints’ are not shown in the paper (similar to my first main point). Given such an interesting dataset, I would expect the authors to deepen their analysis here, by looking at flux and atmospheric CO₂ differences in more spatiotemporal detail.

L626: The difference between coarse and fine grids should be tested in WRF, for example by only transporting emissions from power plants at different resolutions. I do think a difference between coarser and finer resolutions will exist, but the longer-term (yearly) averaging and mid-day selection will counter a large part of this.

L629: This makes me wonder why the authors have not tested a ‘more realistic vertical allocation’.

L633: ‘This study shows where inventories diverge within a megacity’. I am not sure what the authors refer to here. Showing where inventories diverge was already done in Fig. 4, and is not the innovation the authors have achieved in this manuscript.

Abstract:

- Mention the 1km resolution of WRF
- ‘The simulations reproduce most observed ΔCO_2 variability’ I’m not convinced by this, as $R^2 < 0.5$